Abstract Title:
A Systems Biology Knowledgebase and Analysis Platform for *De Novo* Phenotype Predictions, Integrated Omics Analysis, and Iterative Model Improvement.

Authors and Affiliations:
Edward J. O'Brien[1]*, Joshua A. Lerman[1], Roger L. Chang[1], Daniel R. Hyduke[1], Karsten Zengler[1], Bernhard Ø. Palsson[1], **Michael A. Saunders[2]**

[1]Department of Bioengineering, University of California—San Diego, La Jolla, CA, U.S.A.,
[2]Department of Management Science and Engineering, Stanford University, Stanford, CA, U.S.A.
*Presenter.

Presenter contact: ejobrien@ucsd.edu
URL: http://systemsbiology.ucsd.edu

Project Goals:

This project aims to: (1) reconstruct and refine genome-scale models of macromolecular synthesis and metabolism for *Thermotoga maritima* and *Escherichia coli* (2) develop modeling conventions and simulation procedures for phenotype prediction with the integrated models, (3) guide the development of algorithms capable of finding optimal steady-state solutions despite the large, sparse, and ill-scaled constraint matrices, (4) use these prototype models to design a software platform and standard operating procedures to reconstruct general multi-subsystem stoichiometric models, (5) develop and enable methods to parameterize and constrain the network with biophysical models and the direct mapping of diverse omics data, and (6) classify the failure modes of the model to prioritize subsystems and regulatory circuits for model expansion.

Project Description:

Over the past decade, the process of reconstructing metabolic networks at the genome scale has become prevalent in molecular systems biology. There is growing interest in using these metabolic models for both *de novo* phenotype prediction and analysis of omics datasets. To this end, constraint-based algorithms and model-driven omics analysis have provided insight into gene regulation, adaptive evolution, microbial communities, metabolic engineering, and drug response and design. While metabolic models have proven to be a powerful tool, the genetic content of these models is formalized by a Boolean mapping between genetic loci and metabolic reactions, termed the gene-protein-reaction (GPR) relationship. This modeling paradigm only allows for heuristic analysis of transcriptomic and proteomic data and Boolean predictions of genetic requirements.

We have previously shown that it is possible to construct a genome-scale model of RNA and protein expression based on a set of basic biochemical reactions. This process was first completed in *Escherichia coli,* and the resulting model was called the 'E'-matrix (which stands for gene Expression). An analogous macromolecular synthesis reconstruction has

also been completed for *Thermotoga maritima*, which relied heavily on experimental refinement of the transcription unit architecture. The metabolic and the macromolecular synthesis networks have subsequently been merged into integrated models (termed the 'ME' matrix, for Metabolism and Expression), which allow for explicit analysis and simulation of transcriptomes and proteomes in the context of the underlying reaction network. Not only does inclusion of gene expression increase the scope of the model, but the interdependency of gene expression and metabolism also constrains and refines the metabolic solution space, leading to more accurate predictions.

The ME model formulation additionally leads to a reduced dependence on artificial objective functions, such as the biomass objective function, which do not have a mechanistic biochemical basis. For example, nucleotides and amino acids are no longer drawn out of the cell in bulk; instead, individual RNA and protein synthesis fluxes are decision variables in the optimization problem and the *in silico* cell must decide how to invest its finite resources to synthesize them. This framework is shown to capture known trends in the cellular composition of RNA and protein at various growth rates. Allowing for variable cell composition makes the model more generally applicable to diverse environments whereas the biomass objective function tunes the model to a particular condition.

Our experience reconstructing and analyzing these prototype models made it clear that future model refinement, expansion to other subsystems, omics analysis, and reconstructions for other organisms requires a new software platform and standard operating procedures (SOPs). Analogous software and SOPs exist for metabolic models, but certain features of multi-subsystem models necessitate a redesign of the reconstruction framework including: 1) an order of magnitude larger number of reactants and reactions, 2) many different types of molecules (e.g. rna, proteins, metabolites), 3) 'template' reactions for common cellular processes, 4) efficiency parameters coupling different subsystems, 5) the need for direct integration with growing omics datasets, and 6) future expansions to include other cellular processes and subsystems. Due to these challenges and opportunities, we have developed a database-driven solution to reconstruct, query, and generate these multi-subsystem models. We have designed this software to directly couple with model simulation and analysis software, and have included features and flexibility to address the points listed above.

Future efforts will focus on refining model parameters and constraining condition-specific flux variables with omics datasets and biophysical models. Model improvement will then be mediated by the identification of failure modes to prioritize scope expansions (including signaling and regulatory interactions). We demonstrate the promise for long-term applications of this type of model for metabolic and protein engineering, interpretation of adaptive evolution, and analysis of cellular regulation and optimality.

Grant Information:

Numerical Optimization Algorithms and Software for Systems Biology (DOE Award DE-SC0002009)